IFAU – INSTITUTE FOR LABOUR MARKET POLICY EVALUATION

Matching estimators for the effect of a treatment on survival times

Xavier de Luna
Per Johansson

WORKING PAPER 2007:1

The Institute for Labour Market Policy Evaluation (IFAU) is a research institute under the Swedish Ministry of Employment, situated in Uppsala. IFAU's objective is to promote, support and carry out: evaluations of the effects of labour market policies, studies of the functioning of the labour market and evaluations of the labour market effects of measures within the educational system. Besides research, IFAU also works on: spreading knowledge about the activities of the institute through publications, seminars, courses, workshops and conferences; influencing the collection of data and making data easily available to researchers all over the country.

IFAU also provides funding for research projects within its areas of interest. The deadline for applications is October 1 each year. Since the researchers at IFAU are mainly economists, researchers from other disciplines are encouraged to apply for funding.

IFAU is run by a Director-General. The authority has a board, consisting of a chairman, the Director-General and seven other members. The tasks of the board are, among other things, to make decisions about external grants and give its views on the activities at IFAU. A reference group including representatives for employers and employees as well as the ministries and authorities concerned is also connected to the institute.

Postal address: P.O. Box 513, 751 20 Uppsala
Visiting address: Kyrkogårdsgatan 6, Uppsala
Phone: +46 18 471 70 70
Fax: +46 18 471 70 71
ifau@ifau.uu.se
www.ifau.se

Papers published in the Working Paper Series should, according to the IFAU policy, have been discussed at seminars held at IFAU and at least one other academic forum, and have been read by one external and one internal referee. They need not, however, have undergone the standard scrutiny for publication in a scientific journal. The purpose of the Working Paper Series is to provide a factual basis for public policy and the public policy discussion.

ISSN 1651-1166

Matching estimators for the effect of a treatment on survival times*

Xavier de Luna† and Per Johansson‡

16th January 2007

Abstract

We perform inference on the effect of a treatment on survival times in studies where the treatment assignment is not randomized and the assignment time is not known in advance. We estimate survival functions on a treated and a control group which are made comparable through matching on observed covariates. The inference is performed by conditioning on waiting time to treatment, that is time between the entrance in the study and treatment. This can be done only when sufficient data is available. In other cases, averaging over waiting times is a possibility, although the classical interpretation of the estimated survival functions is lost unless hazards are not functions of the waiting times. To show unbiasedness and to obtain an estimator of the variance, we build on the potential outcome framework, which was introduced by J. Neyman in the context of randomized experiments, and adapted to observational studies by D. B. Rubin. Our approach does not make parametric or distributional assumptions. In particular, we do not assume proportionality of the hazards compared. Small sample performance of the estimator and a derived test of no treatment effect are studied in a Monte Carlo study.

*We are grateful to Göran Broström, Peter Frederiksson, Mette Harhoff, Niels Keiding and Michael Rosholm for helpful comments. Comments from seminar participants at the RTN-workshop (Madrid, November 2006) are also acknowledged. The work reported in this article was financially supported by the Institute for Labour Market Policy Evaluation and the Swedish Council for Working Life and Social Research.

†Department of Statistics, Umeå University, 90187 Umeå, Sweden, xavier.deluna@stat.umu.se

‡Institute for Labour Market Policy Evaluation and Department of Economics, Uppsala University, 75120 Uppsala, Sweden, per.johansson@ifau.uu.se

1 Introduction

In order to illustrate the type of studies we address in this paper, let us consider the Stanford heart transplant data set previously analyzed by, e.g., Crowley and Hu (1977) and Kalbfleisch and Prentice (1980). The data set consists in survival times of potential heart transplant recipients after their acceptance into the Stanford heart transplant program. The choice of heart recipients is not randomized in the program. In such an observational study background characteristics affecting both treatment assignment and survival time must be controlled for when evaluating the effect of heart transplantation. One of the peculiarities of the Stanford program, which complicates the analysis (see Keiding, 1995), is that individuals may change treatment status during the follow-up time, being first controls (not transplanted) and later treated (transplanted). Thus, survival times are censored due both to external reasons (e.g., end of study, drop out) and to internal reasons (treatment).

In this paper, we show how a treatment effect can be estimated non-parametricaly in studies such as the Stanford program. We propose to estimate survival functions on a treated and a control group which are made comparable through matching on observed covariates. The inference is performed by conditioning on waiting time to treatment, that is time between the entrance in the study and treatment. This can be done only when sufficient data is available. In other cases, averaging over waiting times is a possibility, although the classical interpretation of the estimated survival functions is lost unless hazards are not functions of the waiting times. To justify the estimator and the related inference, we build on the potential outcome framework, which was introduced by Neyman (1990, translation of a text published in 1923) in the context of randomized experiments, and adapted to observational studies by Rubin (1974); see also Holland (1986) for a review. We work entirely within a discrete time framework, which is consistent with how the data is observed. Our approach does not make any parametric or distributional assumptions. In particular, we do not assume proportionality of the hazards compared, which is equivalent to a constant multiplicative treatment effect. Note that Heller and Venkatraman (2004) recently proposed a non-parametric test of the no-treatment effect hypothesis. Their proposal does, however, not allow for the actual estimation of a treatment effect.

Parametric models have been treated in the literature in connection

with the estimation of treatment effects, see, e.g., Robins (1999), Hernán, Brumback and Robins (2001) and Abbring and van der Berg (2003). The literature on parametric modelling was discussed in Fredriksson and Johansson (2004), who also cast some of the ideas developed and studied herein.

Observational studies of the Stanford program type may be found in other fields than medical applications, including labor economics, where the interest often lies on the estimation of the effect of a training program (treatment) on, e.g., unemployment duration. Such a study is presented in Section 5. While we have found the Stanford heart transplant study to be appropriate as a red thread to illustrate the concepts and methods developed in the paper because of its simplicity and previous use in the literature, the case study of Section 5 is a more realistic application, because of the richness of background information on the individuals. Moreover, the large number of individuals in the study allows us to conduct inference conditionally on waiting time.

The remainder of the paper is organized as follows. The next section describes the inferential issues of interest on an intuitive level. Section 3 presents the potential outcome framework and its associated Neyman's inference, and generalizes it to survival time outcome. For non-censored real valued outcomes this approach has a long history for observational studies, see, e.g., Cochran and Rubin (1973), Rubin (1973a, 1973b, 1990b), and Rosenbaum and Rubin (1984, 1985). In Section 4 the treatment effects of interest are defined. These are basically differences in hazard and survival functions for the treated and the controls. These functions are made comparable through matching on observed covariates while conditioning on waiting time to treatment. Thus, matching estimators are proposed. They are shown to be unbiased and an estimator of their variance is deduced. In Section 5, we use a large data-set on Swedish unemployed and estimate the effect of an employment subsidy on their unemployment duration. In Section 6, a Monte Carlo study is conducted to analyze the finite sample performance of the proposed estimators and a corresponding test of no treatment effect. Finally, the paper is concluded in Section 7.

2 Matching treated with controls

We, purposely, begin by presenting the inferential issues on an intuitive level and delay its formal justification to Section 3 and 4 in order to im-

prove the readability of the paper. We use the Stanford study described in the introduction as background for the discussion. The survival times of the individuals in the study are schematically displayed using Lexis diagrams in Figure 1. In Panel A of the figure, survival times of treated individuals, i.e., patients having received a heart transplant, are represented. Survival times without treatment, i.e. for control individuals, are displayed in Panel B. Note that individuals of Panel A are also included in Panel B since treated patients are not treated until they obtain treatment. Treatment assignment is not made at entrance in the study but only when a heart becomes available. Hence, deleting treated patients from the control group implies a conditioning on survival outcomes, thereby implying a biased analysis.

In both panels, the x-axis represents calendar time with the origin at the beginning of the study, while the y-axis represents time with origin at the time patients are treated in Panel A, and at the time patients enter the study in Panel B**.** Treatment assignments are represented by an open circle in both panels. The diagonal lines represent the history of each patient, starting with their entrance in the study and finishing with death (denoted with a filled circle) or censoring. The solid part of the line highlights the survival time (outcome) of interest, while the dashed part of the line is part of the survival history, however, not part of the outcome.

We want to know which controls can be compared to which treated, or, in other words, we want to select a control group which can be compared to the treated patients, in an evaluation of the effect of the treatment on the survival times (solid diagonal lines).

Randomized treatment: Assume first that treatment is randomized, that is each time a heart transplant can be performed a patient is randomly chosen for treatment from those still alive. Then, in contrast with usual randomized studies, the treated and controls cannot directly be compared, because on average the observed survival durations after transplatation are shorter than the survival durations for the controls. This problem can be corrected by conditioning inference on waiting time. For that purpose, you consider all patients having been transplanted after a given waiting time W, and use as a control group all those that have survived and have not been treated until time W. This is illustrated in the Lexis diagrams of Figure 1 for the first patient to receive a transplantation in Panel A. She/he has waited time W before being treated and, therefore, controls for this treated patient are all those patients passing the horizontal dotted-

dashed line alive and untreated in Panel B.

Observed treatment: Randomization of the treatment is seldom possible for ethical and/or practical reasons. In the Stanford heart transplant program there was no randomization of treatment and, therefore, one should not only condition for the waiting time W but also match for other pre-treatment characteristics. In this context, the crucial assumption is that conditional on waiting time and the observed pre-treatment characteristics affecting both the treatment and the outcome, the treatment assignment can be considered as randomized-like, see Rosenbaum and Rubin (1983). This assumption is sometimes called unconfoundedness assumption. Thus, for a patient having been transplanted, a control having survived until time W without being treated is selected such that her/his pre-treatment characteristics are similar to the treated. This is a so called matching procedure to construct a relevant control group; see, e.g., Rubin (1973a,1974).

Estimation of a treatment effect: When a control group is constructed by a matching procedure −we call such a group a matched control group in the sequel−, it remains to compare survival times for the treated with the survival times for the controls. Average survival times cannot be estimated because censoring arises in both groups. On the other hand, hazards of death and survival functions can be computed and compared for both groups. An extra difficulty arises in the control group since patient's survival may not only be censored by an external mechanism (such as the end of the study) but also by the fact that some controls receive treatment. However, we show in the next section that, under certain conditions, the unconfoundedness assumption yields that the censoring due to treatment is independent of the outcome when conditioning on pre-treatment characteristics.

3 Inferential framework

3.1 Potential outcome specification

Potential outcomes were introduced by Neyman (1990) in the context of randomized experiments as a framework to perform inference on treatment effects. Rubin (1974) generalized their applicability to the context of observational studies. We adapt below the potential outcome framework to the context described in the previous section.

Table 1: Observed status of some variables for a subset of patients among those alive at time $W = 21$ and not treated before that time. We use the convention that for a given day death precedes always treatment, and death preceeds always censoring.

patient ident.	$D(21)$	$T^1(21)$	$T^0(21)$
101	0	NA	C@10
66	0	NA	21
4	0	NA	T@15
47	1	51	NA
97	1	C@110	NA
58	1	321	NA

Note: NA for non-available; C@t for censored at time t after W; T@t for treated at time t.

For an individual which, up to time W, has both survived and has not been treated, we define two potential outcomes:

$$\begin{aligned} T^1(W) &= \text{survival time after time } W \text{ if treated at } W, \\ T^0(W) &= \text{survival time after } W \text{ if neither treated at } W \text{ nor later.} \end{aligned}$$

We, further, denote by $\mathbf{X}$ the vector of pre-treatment characteristics, and by

$$D(W) = \begin{cases} 1 \text{ if treated at time } W, \\ 0 \text{ if not treated at time } W, \end{cases}$$

the treatment indicator. While, in general, $\mathbf{X}$ and $D(W)$ are observed, at most one of the two potential outcomes $T^0(W)$ or $T^1(W)$ can be observed for a given individual having survived until time W. As an example, we summarize in Table 1 the observed status of the variables for a subset of individuals from the Stanford heart transplant study given the waiting time $W = 21$.

We make a first assumption, often called the stable-unit-treatment-value assumption; see, e.g., Rubin (1990b).

Assumption A: *The values* $T^1(W)$ and $T^0(W)$ *for a given individual are not affected by the values taken by* $D(W)$ *for any other individual.*

3.2 Inference

Assume now that at a given time W in a study, n_1 individuals are treated, indexed by $i = 1, 2, \ldots, n_1$. A certain amount of individuals (often $\gg n_1$) have also survived until time W although they are not treated at that time, thereby providing a reservoir of controls. A matched control group is extracted from this reservoir as described in Section 2, indexed by $i = n_1+1, \ldots, 2n_1$. In this paper, we focus on estimands of the following type:

$$\Delta(W) = \frac{1}{2n_1} \sum_{i=1}^{2n_1} \left(T_i^1(W) - T_i^0(W) \right),$$

i.e. the average treatment effect for treated patients and their match in the control group for a given waiting time. Other type of estimands based on different populations may be considered, see Rubin (1991) and Imbens (2004). For instance, one may be interested in "the average treatment effect for a future patient exposed to treatment, where treatment is assigned with $p(\mathbf{X}, W)$ at a given time W". These two types of estimands are called "average treatment effect on the treated" in the econometrics literature, see, e.g., Imbens (2004). The former, $\Delta(W)$, is a sample average, with missing values, while the latter is defined on a super-population. The information available in the observed sample on $\Delta(W)$ is also all the information we have on the estimand defined on the super-population. Moreover, for an unbiased estimator of $\Delta(W)$ to be an unbiased estimator of the second estimand, we need to further assume that the patients in the study are representative (e.g., a random sample) of the super-population implicitely defined. Yet another estimand of interest might be "the average treatment effect for a future patient on which treatment is imposed at a given waiting time W". This is again a super-population estimand and, in general, to obtain an unbiased estimator one would need to match both the treated with similar controls and the controls with similar treated. All the estimands described are equivalent if we have *constant additive treatment effect*, i.e. $T^1(W) = T^0(W) = \delta$ for all individuals.

To perform inference, we need a probability model. Several models and resulting modes of inference may be entertained, see Rubin (1990b, 1991). In any case, a model for the treatment assignment mechanism is the corner stone to the identifiability of the estimand of interest. That is a specification of $\Pr(D(W) = 1 | \mathbf{X}, T^1(W), T^0(W))$ the probability of being

treated when the waiting time has been W. The following two assumptions are often made within the potential outcome framework.

Assumption B: *The assignment mechanism $D(W)$ is independent of the potential outcomes, conditional on the set of pre-treatment characteristics* $\mathbf{X}$, *i.e.,* $\Pr(D(W) = 1|\mathbf{X}, T^1(W), T^0(W)) = p(\mathbf{X}, W)$, *where $p(\mathbf{X}, W)$ is a function of $(\mathbf{X}, W)$ only; see Dawid (1979) for an account on conditional independence statements.*

Assumption C: $0 < p(\mathbf{X}, W) < 1$.

Assumption B is sometimes called the unconfoundedness assumption. Assumption C states that each individual having survived without being treated until time W has non-zero probability of being both treated and not treated at time W. Assumption B and C together were termed strong ignorability of the treatment assignment by Rosenbaum and Rubin (1983). They guarantee that a treated and a control individual having the same value for $\mathbf{X}$ (a matched pair) can be compared in order to infer a treatment effect.

If we would have uncensored survival times, an estimator of $\Delta(W)$ would be

$$\widehat{\Delta}(W) = \frac{1}{n_1} \sum_{i=1}^{n_1} \left(T_i^1(W) - T_{i+n_1}^0(W) \right), \tag{1}$$

where we assume that individual $i + n_1$ is a match to individual i, $i = 1, \ldots, n_1$. The properties of the Statistic (1) can be studied by considering its sampling distribution under treatment reassignments through $p(\mathbf{X}, W)$ for fixed values of $T_i^1(W), T_i^0(W)$, $i = 1, \ldots, 2n_1$, with the constraint that within each matched pair both treatment and non-treatment arise. Under this assignment mechanism, over the $\binom{2n_1}{n_1}$ randomizations, we have unbiasedness (Neyman, 1990, Rubin, 1990b): $E(\widehat{\Delta}(W)) = \Delta(W)$. If, moreover, we have constant additive treatment effect, then

$$\frac{1}{n_1} \sum_{i=1}^{n_1} \left((T_i^1(W) - T_{i+n_1}^0(W)) - \widehat{\Delta}(W) \right)^2 \tag{2}$$

is an unbiased estimate of the variance of $\widehat{\Delta}(W)$. This mode of inference dates back to Neyman (1990) and is called by Rubin (1990a) randomization-based repeated-sampling inference; see also the Appendix.

4 Survival analysis

4.1 Hazards: estimand and estimators

The estimator $\widehat{\Delta}(W)$ above cannot be computed, because the survival times of the observed individuals are censored. Controls can be censored by treatment and both controls and treated can be censored, for instance, by the end of the study period. In such situations, it is customary to use the survival analysis approach, see, e.g., Kalbfleish and Prentice (1980). Assumptions on the censoring mechanisms must be made, however.

First, we want censoring due to treatment after time W to be independent of the potential outcome $T^0(W)$ conditional on $\mathbf{X}$. To shed light on this issue let $C^T(W)$ denote the time to treatment for an individual not treated at time W. By convention $T^0(W)$ is censored when $C^T(W) < T^0(W)$.We have, for $j < t^0$,

$$\begin{aligned}&\Pr(C^T(W) = j|\mathbf{X}, T^0(W) = t^0)\\ &= \Pr(D(W+j) = 1, D(W+l) = 0, l = 1, \dots j-1|\mathbf{X}, T^0(W) = t^0).\end{aligned}$$

In order to be able to take advantage of Assumption B we assume that the following decomposition holds.

Assumption D: *For* $j < t^0$,

$$\begin{aligned}&\Pr(D(W+j) = 1, D(W+l) = 0, l = 1, \dots j-1|\mathbf{X}, T^0(W) = t^0)\\ &= \Pr(D(W+j) = 1|\mathbf{X}, T^0(W) = t^0) \Pr(D(W+j-1) = 0|\mathbf{X}, T^0(W) = t^0)\\ &\qquad \times \cdots \times \Pr(D(W+1) = 0|\mathbf{X}, T^0(W) = t^0).\end{aligned}$$

Assumption D says that treatment assignments at different times (after time W) are independent of previous assignment when conditioning on $\mathbf{X}$ and $T^0(W)$. Under this assumption and Assumption B, we can then write

$$\begin{aligned}&\Pr(C^T(W) = j|\mathbf{X}, T^0(W) = t^0)\\ &= \Pr(D(W+j) = 1|\mathbf{X}) \Pr(D(W+j-1) = 0|\mathbf{X}) \cdots \Pr(D(W+1) = 0|\mathbf{X}).\end{aligned} \tag{3}$$

By the latter equality we see that with Assumptions B and D we obtain a censoring mechanism due to treatment which is independent of the potential outcome $T^0(W)$ when conditioning on $\mathbf{X}$.

Let us now define the variable $C^E(W)$, the time to censoring (by other reasons than treatment, e.g. end of study and drop out) for an individual having survived until time W. Then, the observed survival time is censored depending on whether $C^E(W) < T^1(W)$, or $C^E(W) < T^0(W)$. We make the following assumption.

Assumption E: $C^E(W)$ *is independent of* $T^1(W)$ *and of* $T^0(W)$ *when conditioning on* $\mathbf{X}$.

Assumption E corresponds to usual hypotheses of independent censoring mechanism made in survival analysis. We restrain here to introduce a new notation to denote censored potential outcomes. Thus, in the sequel, $T^k(W)$, $k = 0, 1$, denotes time to death or to censoring.

In a survival analysis approach, instead of comparing the average survival times, $\Delta(W)$ above, the sample hazards (proportion of individuals dying at time t among those having survived up to time t) are compared for the treated and controls. We therefore consider the estimand

$$\Delta_h(t; W) = h^1(t; W) - h^0(t; W), \tag{4}$$

where

$$h^k(t; W) = \frac{\sum_{i=1}^{2n_1} I(T_i^k(W) = t)}{\sum_{i=1}^{2n_1} I(T_i^k(W) \geq t)}, \text{ for } k = 0, 1.,$$

where $I(T \geq t) = 1$ if $T \geq t$, i.e. if the individual has survived and is not censored until time t, and $I(T \geq t) = 0$ otherwise. Also, $I(T = t) = 1$ if $T = t$ because of death (not censoring) and $I(T = t) = 0$ otherwise.

Estimand $\Delta_h(t; W)$ and its building blocks $h^k(t; W)$, $k = 0, 1$, are defined with respect to the potential outcomes which have been censored. Another estimand could have been defined based on the uncensored potential outcomes. Our focus on the censored version is justified by Assumption E and by (3), consequence of Assumptions B and D, which together guarantee that the $\Delta_h(t; W)$ defined under censoring is representative (in a frequentist sense) of the same estimand without censoring. Note, moreover, that under a zero constant additive treatment effect, $T_i^1(W) - T_i^0(W) = 0$, for all i, the estimands with and without censoring are equivalent.

Estimand $\Delta_h(t; W)$ can be estimated with

$$\widehat{\Delta}_h(t; W) = \widehat{h}^1(t; W) - \widehat{h}^0(t; W),$$

where

$$\widehat{h}^k(t; W) = \frac{\sum_{i=1}^{n_1} I(T^k_{ik+(i+n_1)(1-k)}(W) = t)}{\sum_{i=1}^{n_1} I(T^k_{ik+(i+n_1)(1-k)}(W) \geq t)}, \text{ for } k = 0, 1.$$

We show in the Appendix that, under the Assumptions A-C, $\widehat{\Delta}_h(t;W)$ is unbiased for $\Delta_h(t;W)$: $E\left(\widehat{\Delta}_h(t;W)\right) = \Delta_h(t;W)$. Moreover, the variance of the estimator $\widehat{\Delta}_h(t;W)$ can be estimated with

$$\widehat{Var}\left(\widehat{\Delta}_h(t;W)\right) = \frac{\widehat{h}^1(t;W)(1-\widehat{h}^1(t;W))}{\sum_{i=1}^{n_1} I(T_i^1(W) \geq t) - 1} + \frac{\widehat{h}^0(t;W)(1-\widehat{h}^0(t;W))}{\sum_{i=1}^{n_1} I(T_{i+n_1}^0(W) \geq t) - 1}.$$

This estimator is unbiased, for instance, when there is no treatment effect in the sense that $T_i^1(W) = T_i^0(W)$, for $i = 1, \ldots, 2n_1$. In general, however, it is positively biased (yielding conservative inference); see the Appendix. This is a qualitatively different result from that of Neyman (1990), where unbiasedness of the variance estimator (2) is guaranteed under constant additive treatment effect. This difference is due to the fact that the hazard is based on indicator functions of the survival times and not on the times themselves.

4.2 Survival function

The survival function, the proportion of individuals surviving at least up to time t, constitutes a convenient way to summarize or aggregate the information from the hazards calculated above. Denote by $T_{(1)}^1(W) \leq T_{(2)}^1(W) \leq \cdots \leq T_{(m_1)}^1(W)$ the $m_1 \leq 2n_1$ not censored survival times if treated, sorted in ascendant order. Then, the survival function when treated is defined as

$$F^1(t;W) = \prod_{i:T_{(i)}^1 < t} (1 - h^1(T_{(i)}^1(W);W)). \tag{5}$$

Similarly, we define the survival function when not treated by

$$F^0(t;W) = \prod_{i:T_{(i)}^0 < t} (1 - h^0(T_{(i)}^0(W);W)), \tag{6}$$

where $T_{(1)}^0(W) \leq T_{(2)}^0(W) \leq \cdots \leq T_{(m_0)}^0(W)$ are the $m_0 \leq 2n_1$ not censored survival times if not treated, sorted in ascendant order. We are, thus, interested in estimating the difference

$$\Delta_s(t;W) = F^1(t;W) - F^0(t;W). \tag{7}$$

An estimator of $\Delta_s(t;W)$ is readily available by replacing the hazards by their estimators described above, as follows

$$\widehat{F}^1(t;W) = \prod_{i:\widetilde{T}^1_{(i)}<t} (1 - \widehat{h}^1(\widetilde{T}^1_{(i)}(W);W)),$$

where $\widetilde{T}^1_{(1)}(W) \leq \widetilde{T}^1_{(2)}(W) \leq \cdots \leq \widetilde{T}^1_{(\widetilde{m}_1)}(W)$ are the $\widetilde{m}_1 \leq n_1$ observed and not censored survival times for the treated individuals, and

$$\widehat{F}^0(t;W) = \prod_{i:\widetilde{T}^0_{(i)}<t} (1 - \widehat{h}^0(\widetilde{T}^0_{(i)}(W);W)),$$

where $\widetilde{T}^0_{(1)}(W) \leq \widetilde{T}^0_{(2)}(W) \leq \cdots \leq \widetilde{T}^0_{(\widetilde{m}_0)}(W)$ are the $\widetilde{m}_0 \leq n_1$ observed and not censored survival times for the matched control individuals. These estimators of the survival functions are usual Kaplan-Meier estimators (Kaplan and Meier, 1958).

The asymptotic variance of the estimated survival functions can be estimated by, for $j = 0, 1$,

$$\widehat{Var}(\widehat{F}^j(t;W)) = \left[\widehat{F}^j(t;W)\right]^2 \times \sum_{i:\widetilde{T}^j_{(i)}<t} \frac{\widehat{h}^j(\widetilde{T}^j_{(i)}(W);W)}{\sum_{k=1}^{n_1} I(T^j_{kj+(k+n_1)(1-j)}(W) \geq \widetilde{T}^j_{(i)}(W)) - \sum_{k=1}^{n_1} I(T^j_{kj+(k+n_1)(1-j)}(W) = \widetilde{T}^j_{(i)}(W))},$$

see Kaplan and Meier (1958). The above expression is called the Greenwood's formula (Greenwood, 1926). Finally, the estimator based on this asymptotic approximation $\widehat{Var}(\widehat{\Delta}_s(t;W)) = \widehat{Var}(\widehat{F}^1(t;W)) + \widehat{Var}(\widehat{F}^0(t;W))$ is, as in Section 4.1, expected to be conservative when the treatment effect is not exactly zero for all individuals. The simulation study of Section 6 shows that this estimated variance can be useful to test the hypothesis of no treatment effect.

4.3 Averaging over waiting times

The theory above has been developed for a fixed waiting time to treatment, W. However, estimating the survival functions non-parametrically

for a given waiting time assumes the availability of sufficiently many observations at each waiting time W of interest. This is not always the case as, e.g., with the Stanford heart transplant data. In such cases, one may average over the observed waiting times yielding the new estimand

$$\overline{\Delta}_h(t) = \overline{h}^1(t) - \overline{h}^0(t), \tag{8}$$

where $\overline{h}^k(t) = \frac{\sum_{i=1}^{2N_1} I(T_i^k = t)}{\sum_{i=1}^{2N_1} I(T_i^k \geq t)}$, for $k = 0, 1$, and now $i = 1, \ldots, N_1$ indexes all the treated individuals in the study, and $i = N_1 + 1, \ldots, 2N_1$ indexes all the corresponding match controls. This is the average treatment effect for treated patients and their match in the control group. The corresponding estimator is obtained by considering all individuals in the treated-matched sample instead of only those with a given waiting time. This estimand and estimator have a clear interpretation (average difference in hazard). However, $\overline{\Delta}_s(t)$ obtained by "plugging in" $\overline{h}^k(t)$, $k = 0, 1$, in (6) and (5), is difficult to interpret unless the hazards $h^k(t; W)$, $k = 0, 1$, are not functions of W. On the other hand, interpretability of $\overline{\Delta}_s(t)$ is less of an issue if the main objective of the analysis is to test the hypothesis of no treatment effect.

Finally, computing and estimating the variance of such an estimator is difficult since some treated may be used as control. However, in cases where such double use of the same individual is rare (e.g., many untreated controls are available) the variance provided in the previous section may be used when averaging over waiting times as well.

4.4 The Stanford heart transplant program

We replicate the analysis of Crowley and Hu (1977) based on the Cox proportional hazard model. They model the hazard for patient i with

$$h(t_i) \exp(\delta Z(t_i; W_i) + \mathbf{X}_i' \boldsymbol{\beta}) \tag{9}$$

where $h(t_i)$ is the baseline hazard and $Z(t_i; W_i) = I(t_i \geq W_i)$, the heavy side function. We use exact partial maximum likelihood to estimate the parameters of model (9), where the vector $\mathbf{X}_i$ contains age when eligible and year of acceptance into the study. Information on prior surgery is not included because it was found not significant.

The estimate $\widehat{\overline{\Delta}}_s(t)$ is displayed in Figure 2. The estimation is obtained by one to one matching. That is for each treated a control is chosen by

matching exactly on year of acceptance and waiting time to transplant, while matching on the age when eligible on a nearest neighbor (Euclidian distance) basis. The nearest neighbor must not be more than four years apart. This restricts the sample to 60 treated and matched control individuals. Together with $\widehat{\overline{\Delta}}_s(t)$ we display in Figure 2 the difference in survival functions resulting from the fitted Cox proportional hazard model described above. We do not display the 95% confidence bands to improve readability. They clearly include zero. Since there is a large percent of treated in this study, the variance estimator is, however, not reliable as noted in Section 4.3.

Model (9) assumes proportionality of the hazards with and without treatment. Moreover, both the parametric and the matching estimators must assume that hazard functions do not to depend on W due to the small sample available. This together with the fact that there is little background information on patients make the evaluation of the transplantation effect difficult.

5 Effect of an employment subsidy program

To illustrate the use of the non-parametric estimator with a realistic application we use a data set previously analyzed in Forslund, Johansson and Lindqvist (2004). The interest is to estimate the effect on employment of an employment subsidy program targeted at the long-term unemployed, i.e., individuals at least 25 years-of-age and who had been registered as unemployed at the public employment service (PES) for at least 12 months without interruption. The subsidy amounted to 50 percent of total wage costs and was paid for a maximum period of 6 months. The subsidy was also capped at 350 Swedish crowns per day and could be extended to 12 months in some exceptional cases.

Register data from the Swedish National Labor Market Board is used to evaluate the program. The database contains information on all individuals registering at the PES in Sweden since August 1991, including, age, sex, educational attainment, the individuals' registration date and past job training activities.

The individuals in the data are classified into two different groups: those who start the employment subsidy program after having become eligible and eligibles who do not start the program. The study start on

January 1998. Each time a person becomes eligible, the duration (months) until she or he either finds an employment or becomes right censored (end of study on October 1 2002, drop out) is recoded. A total of 631,358 individuals, aged 25–63, were eligible for the program during the study period. Three percent of the eligible spells ended into the program. The most salient feature of the eligible persons is that they, on average, had a long lasting relationship with the employment service; see Forslund et al. (2004) for more details.

To obtain $\widehat{\Delta}_s(t;W)$ and $\widehat{\overline{\Delta}}_s(t)$ we use one to one exact matching. For each participant we look for one control (nonparticipant) which has exactly the same values for a set of covariates: sex, Nordic citizenship, unemployment insurance, disabled, high school degree, university degree−all binary−, age (≤ 30, 31-40, 40-50), number of previous days in unemployment register (0, 1-100, 101-500, 501-1000, >1000), number of previous spells in unemployment register (0, 1-5,6-15), and the local labor market of the individuals (Sweden is divided into 100 local labor markets). All these covariates are expected to affect both the unemployment duration and the participation into the program. The matching estimator is based on $7,651$ treated individuals, thus $12,300$ people in the employment subsidy program are removed due to lack of common support (no matching individual found in the control group).

This study does not suffer of the limitations of the Stanford heart transplant program. Because we have many observations $(7,651)$ we are able to estimate treatment effects by conditioning on W. Moreover, we have few treated (3%) and, therefore, only 265 individuals in the matched control group also belongs to the treated group. This should allow us to use the variance calculations of Section 4.2 for the estimator $\widehat{\overline{\Delta}}_s(t)$. Note that here right censoring concerns 51% of the sample. There are also some drop-outs, see Forslund et al. (2004) for more details on those.

Thus, Figure 3 shows the estimated treatment effects $\widehat{\overline{\Delta}}_s(t)$ and the estimated treatment effects $\widehat{\Delta}_s(t;W)$ for $W = 1, 11, 21$ and, 31. In all cases 95% confidence bands (based on the normal approximation) are also displayed to judge significance.

The conditional (on waiting time) results are similar which indicates that the hazards are fairly constant with respect to waiting time. This enables us to focus our discussion on the average estimate. We see that after an initial period of about 6 months with a negligible (negative) pro-

gram effect there is an downward jump; from then on the effect gradually becomes smaller, but it is positive (i.e., the program shorten unemployment duration) and significant over the rest of the follow-up horizon (57 months). This scenario is consistent with an initial period of locking in effect (i.e. individuals do not find a job while being in the program −they are supposed to seek non-subsidized employment while into the program) and a subsequent period with a positive program effect. The sum of the effects over the whole follow-up horizon is 7.8, which corresponds to a decrease in unemployment duration (from the entrance into the program) by almost 8 months.

As a way of comparison, we tried to fit (using partial maximum likelihood) a Cox regression model separately for the participants, $h^0(t, \mathbf{X}, W) = h^0(t, W)\exp(\mathbf{X}'\beta_0)$, and the non-participants, $h^1(t, \mathbf{X}, W) = h^1(t, W)\exp(\mathbf{X}'\beta_1)$. Using two different Cox regression specification allows us to avoid the assumption of proportionality (with respect to the treatment assignment); see Kalbfleish and Prentice (1980, pp. 136). However, due to the large amount of covariates in comparison with the small amount of participants, this approach was not practically feasible. Paradoxically, the parametric approach is here more data demanding than matching. On the other hand, we can use Rosenbaum and Rubin (1983) result saying that it is sufficient to control for the probability of being treated given $\mathbf{X}$, $p(\mathbf{X}, W)$ if Assumptions B and C hold. Hence, we start by fitting (maximum likelihood) a logistic regression model to obtain fitted values $\widehat{p}(\mathbf{X}, W)$. The covariates used are those matched for earlier, with age, number of previous days unemployed and number of previous spells in unemployment register included as polynomial of order two as well. The controls used are those which have survived at least until time W. Next, we estimate the two Cox regressions $h^0(t, \mathbf{X}, W) = h^0(t, W)\exp(\beta_0\widehat{p}(\mathbf{X}, W))$ and $h^1(t, \mathbf{X}, W) = h^1(t, W)\exp(\beta_1\widehat{p}(\mathbf{X}, W))$. The resulting difference in survival functions are also displayed in Figure 3. In accordance with Assumption C we have used only the controls which have probability $\widehat{p}(\mathbf{X}, W)$ of entering the program as large as the lowest fitted probability for the observed participants. Thus, for instance, 3,005 controls are deleted when not conditioning on waiting time. The results based on the Cox regression on $\widehat{p}(\mathbf{X}, W)$ have the same general pattern as the matching estimators, though the effect is slightly higher in all cases. Note that we expect the variability of the Cox regression survival functions to be larger because conditioning on $\widehat{p}(\mathbf{X}, W)$ instead of $\mathbf{X}$ (matching estimator) is typically

less accurate. However, we are not aware of any theoretical or applied work on the Cox regression approach implemented here, which constitutes an interesting future topic for research.

6 Monte Carlo study

6.1 Design

We study here the small sample performance of the matching estimators introduced in this paper. To this end we generate geometrically distributed survival times. Without loss of generality we consider a situation where all n individuals simulated have same entry time into the study. The study lasts 50 units of time. From the entry time, time to death T_i, for each individual i, is simulated from a geometric distribution function with probability of success (death) equal to $p_0(X_i)$. The probability distribution of time to death is, thus, given by $\Pr(T_i = t_i|X_i) = p_0(X_i)[1 - p_0(X_i)]^{t_i}$. Similarly, we generate the time to treatment T_i^d, for each individual i, from a geometric distribution function with probability of success (treatment) $p_d(X_i)$. The probability distribution of time to treatment is then given by $\Pr(T_i^d = t_i|X_i) = p_d(X_i)[1 - p_d(X_i)]^{t_i}$. For $p_0(X_i)$ and $p_d(X_i)$ we use a logistic function: $p_0(X_i) = (1 + \exp(-(a_0 + a_1X_i))^{-1}$, and $p_d(x_i) = (1 + \exp(-(b_0 + b_1X_i))^{-1}$. We consider the following situations: $a_0 = -3.0$, $b_0 = -5.5$ and -3.0, and $b_1 = a_1$ are set either to 0 or 1. In the homogenous case (i.e. $b_1 = a_1 = 0$) the death hazard if not treated is $p_0(X_i) = 0.047$ and the hazards into treatment are $p_d(X_i) = 0.0041$ and 0.047 respectively for $b_0 = -6.5$ and -3.0. The proportion of treated was on average equal to 2.9 and 49 per cent in these two simulated situations. These designs were chosen to resemble the situation encountered in the two applications described earlier: the employment subsidy where only 3% were treated and the Stanford heart transplant program where we had 67% of treated individuals. In the employment subsidy treatment the unconditional monthly hazard to death (employment in this application) if not treated is approximately constant, around 0.045. In the Stanford heart transplant program the daily unconditional hazard to death if not treated is decreasing: we obtained an unconditional hazard of 7%, 4%, 3% and 1% based on respectively the first 7 deaths, the first 13 deaths, the first 23 deaths, and on all 30 observed deaths. In order to keep the design

simple and transparent we use a constant unconditional hazard of 0.047 in both simulations.

We further need to simulate a time to death T_i^1 from the time of treatment, for those individuals who are treated, i.e. such that $T_i^d < T_i$. We use again a geometric distribution $\Pr(T_i^1 = t_i|X_i) = p_1(X_i)[1 - p_1(X_i)]^{t_i}$, where $p_1(X_i) = (1+\exp(-(a_0+a_w+a_1X_i))^{-1}$. Finally, the X_i's are generated from a $(0,1)$ uniform distribution, and are fixed in repeated samples.

This simulation design can be related to the potential outcome framework of Section 3 as follows. The treatment assignment mechanism at time W_i, $D_i(W_i)$, is given by

$$\begin{aligned}\Pr(D_i(W_i) = 1|X_i) &= \Pr(T_i > W_i \cap T_i^d = W_i|X_i) \\ &= \Pr(T_i > W_i|X_i)\Pr(T_i^d = W_i|X_i).\end{aligned}$$

For a given individual i, the potential outcomes simulated are: $T_i^0(W) = T_i - W$ for individuals such that $T_i > W_i$ and $T_i^1(W) = T_i^1$ for individuals such that $T_i^d = W$. Thus, while we simulate $T_i^0(W)$ for all individuals having survived until time W, we simulate $T_i^1(W)$ only for those treated at time W. A consequence is that the hazard $h^0(t;W)$ is known while $h^1(t;W)$ is not. We, therefore, choose to use their limit (letting the number of treated n_1 tend to infinity) in probability, $h^k(t;W) \xrightarrow{p} \widetilde{h}^k(t;W)$ as $n_1 \to \infty$, $k = 0, 1$, to assess the quality of the estimators. We have

$$\widetilde{h}^k(t;W) = \int \Pr(T^j(W) = t|T^k(W) \geq t, X_i)dx_i \simeq \frac{1}{n}\sum_{i=1}^{n} p_k(X_i)$$

for n, the number of simulated design points, large enough. In particular, when $a_1 = 0$, $\widetilde{h}^k(t;W) = p_k(X_i) = p_k$. Thus, we use $\widetilde{\Delta}_s(t;W)$ obtained by using $\widetilde{h}^1(t;W)$ and $\widetilde{h}^0(t;W)$ as an approximation of the estimand of interest $\Delta_s(t;W)$. This approximation is used to compute the bias of the estimator $\widehat{\Delta}_s(t;W)$ with 1,000 simulated replicates. This is reasonable because the difference $h^k(t;W) - \widetilde{h}^k(t;W)$, $k = 0, 1$, is zero on average (over the replicates).

In general, the hazard $h^0(t;W)$ may depend on W. This is not the case here due to the choice of the geometric distribution for generating survival times. In this experiment we have, therefore, that the difference in hazards depends on W only through a_w. When the latter is constant

with respect to W, we have $\Delta_s(t;W) = \Delta_s(t)$, but otherwise we can make $\Delta_s(t;W)$ depend on W.

The no-treatment effect situation is obtained by letting $T_i^1(W) = T_i^0(W)$. Situations with a non-zero treatment effect are designed for an homogenous ($a_1 = 0$) and an heterogenous ($a_1 = 1$) case as follows. We let $a_w = \ln((0.047 + (1/(W+23)))/(1 - 0.047 - (1/(W+23))))$. For the homogenous case this yields the treatment effects $\widetilde{\Delta}_h(t;1) = 1/24 \cong 4.2\%$, $\widetilde{\Delta}_h(t;5) = 1/28 \cong 3.6\%$, $\widetilde{\Delta}_h(t;10) = 1/33 \cong 3.0\%$, $\widetilde{\Delta}_h(t;15) = 1/38 \cong 2.6\%$. Figure 4 displays the treatment effects $\widetilde{\Delta}_s(t;W)$ for these four values of W as well as for the average. The figure shows clearly that the treatment effect varies with W, getting smaller as waiting time goes.

6.2 Results

We study the bias of the estimator $\widehat{\Delta}_s(t;W))$, where matching is performed as described in the applications, see Section 2. The continuous covariate x_i is matched using the nearest neighbor with respect to the Euclidian distance. Moreover, the size and the power of the test of no treatment effect ($T_i^1(W) = T_i^0(W)$) with the Wald-test statistic

$$\frac{\widehat{\Delta}_s(t;W)}{\widehat{Var}(\widehat{\Delta}_s(t;W))^{1/2}}$$

is also studied. We perform experiments, where the number of individuals are varied as $n = 500$, 1,500 and 6,000, and the number of replicates is 1,000.

To save space, we restrict the presentation of the results to the setting with covariates (i.e. $a_1 = b_1 = 1$). For the bias study, we show results for the case with a treatment effect (i.e. $a_w \neq 0$). The non-reported cases gave a similar picture to those reported. We start by presenting the case with 2.9% treated. Thereafter, results for the situation with 49% treated are commented. The results are displayed in figures with panels ordered from left to right with respect to sample size and from bottom to top with respect to W, except for the panels on the first row, where the results for the average estimator/test are displayed.

6.2.1 The case with 2.9 percent treated

The bias of the estimator and the size and power of the test of no treatment effect are presented in Figures 5, 6 and 7, respectively. In this setting, the

number of treated equals on average 14.5, 44 and 174 when the sample size equal 500, 1,500 and 6,000, respectively. For $W = 1$ and 15 the corresponding figures are 1.63, 3.37 and 9.13 and 1.18, 1.57 and 4.01, respectively.

The bias (Figure 5) is, as expected, decreasing with sample size n. Considering the relatively small sample sizes of treated for each W the estimator does well. The size of the Wald test is displayed in Figure 6. For sample sizes 1,500 and 2,500 the average estimator have approximately correct size. For the conditional estimator the size is too small for small n and too large when $n = 6,000$ and $W > 5$. Considering the extremely few number of treated for each W these results are perhaps not surprising. When extending the sample size to 12,000 the correct size is well approximated for all values of W. The power of the Wald test is displayed in Figure 7. As expected the Wald test has significant power only for the average effects, due to the low number of treated individuals in each sub-group defined by W.

6.2.2 The case with 49 percent treated

The bias of the estimator, and the size and power of the test of no treatment effect are presented in Figures 8, 9 and 10, respectively. In this setting the number of treated is for sample size of 500, 1,500 and 6,000 on average 240, 718 and 2,876, respectively. For $W = 1$ and 15 the corresponding figures are 30, 90 and 363 and 4, 10 and 39, respectively.

The bias displayed in Figure 8 is positive however decreasing with n. The size of the Wald test is displayed in Figure 9. There is a small tendency of too large size when n is small for the conditional estimator. The average estimator always displays too large a size. Larger sample sizes do not help in this situation. Because the fraction of treated is large and constant, this result should be expected. The power of the Wald test is displayed in Figure 10. It increases with n. The power of the test for the average case is not comparable to the others due to the size failure.

7 Concluding discussion

We have proposed a non-parametric estimator of a treatment effect on a survival outcome. The effect (estimand) is a difference of survival functions

estimated on two groups of matched individuals (treated and control). The estimators and their inference presented are best suited for observational studies including many individuals. Such large observational studies allows us to relax restrictive assumptions such as parametric functional forms, proportionality of the hazards, and homogeneity of the hazards with respect to waiting time until treatment.

To avoid assumptions of constant hazard with respect to waiting time until treatment we propose to perform inference conditionally on the waiting time, when the proportion of treated individuals at each waiting time of interest is large enough. Otherwise, one can average over waiting times. In the latter case, the estimated treatment effect does not keep necessarily its interpretation of a difference of survival functions, although one may still test the hypothesis of no treatment effect. The variance estimator we provide is in general conservative unless there is zero constant additive treatment effect. This allows us to test such a no treatment effect hypothesis. In our simulations, the empirical size of the test was close to the nominal, except when averaging over waiting times in situations where the percentage of treated individuals was large. The test was also shown to have power, although, it is essential to observe enough treated individuals at a given waiting time, when conditioning the inference on the latter.

An application on an employment subsidy to shorten unemployment duration is used to illustrate the applicability of the proposed matching estimator. Although our applications did not include time-dependent covariates, such situations are straighforward to handle since such covariates are not time-dependent anymore when conditioning the inference on waiting time. We also suggest and use an alternative evaluation method to matching, where Cox regression models are fitted on treated and controls separately with $\Pr(D(W) = 1|\mathbf{X})$ as unique covariate. Conditioning solely on this covariate is justified by the results of Rosenbaum and Rubin (1983). However, because $\Pr(D(W) = 1|\mathbf{X})$ is unknown and must be fitted, it is not straightforward how inference on the resulting survival functions should be performed. This is an interesting direction of future research. Note, finally, that the matching estimators introduced herein could also make use of Rosenbaum and Rubin's results, by matching on $\Pr(D(W) = 1|\mathbf{X})$ instead of $\mathbf{X}$. This is often advocated in the matching literature, for instance, when $\mathbf{X}$ includes many continuous covariates in order to diminish the bias due to poor matches.

References

Abbring, J. H. and van der Berg, G. J. (2003), The non-parametric identification of treatment effects in duration models, *Econometrica* **71**, 1491-1518.

Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: A review, *Sankhya Ser. A* **35**, 417-446.

Crowley, J. and Hu M. (1977). Covariance analysis of heart transplant data, *Journal of the American Statistical Association* **72**, 27-36.

Dawid, A. P. (1979). Conditional independence in statistical theory, *Journal of the Royal Statistical Society Ser. B* **41**, $1 - 31$.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review, *The Review of Economics and Statistics* **86**, 4-29.

Forslund, A., Johansson, P., and Lindqvist, L. (2004). Employment subsidies: A fast lane from unemployment to work?, Working Paper 2004:18, Institute for Labour Policy Evaluation, Uppsala.

Fredriksson, P. and Johansson, P. (2004). Dynamic treatment assignment − The consequences for evaluations using observational data, Discussion Paper 1062, Institute for the Study of Labor, Bonn.

Greenwood, M. (1926). The natural duration of cancer, *Reports on Public Health and Medical Subjects* **33**, 1-26, His Majesty's Stationery Office: London.

Heller, G. and Venkatraman, E. S. (2004). A nonparametric test to compare survival distributions with covariance adjustment, *Journal of the Royal Statistical Society Ser. B* **66**, 719-733.

Hernán, M., Brumback, B., and Robins J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments, *Journal of the American Statistical Association* **96**, 440-448.

Holland, P. (1986). Statistics and causal inference, with discussion, *Journal of the American Statistical Association* **81**, $945 - 970$.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data,* Wiley: New York.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457-481.

Keiding, N. (1995). Historical controls and modern survival analysis, *Lifetime Data Analysis* **1**, 19-25.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Translated by D. M. Dabrowska and edited by T. P. Speed, *Statistical Science* **5**, 465-472.

Rosenbaum, P. R and Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika* **70**, $41 - 55$.

Rosenbaum, P. R and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score, *Journal of the American Statistical Association* **79**, 516-524.

Rosenbaum, P. R and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling incorporating the propensity score, *American Statitistician* **39**, 33-38.

Robins, J. M. (1999). Marginal structural models versus structural nested models as tools for causal inference, in *Statistical Models in Epidemiology: The Environment and Clinical Trials*, M.E. Halloran and D. Berry (eds), IMA Volume 116, Springer-Verlag, New York, 95-135.

Rubin, D. B. (1973a). Matching to remove bias in observational studies, *Biometrics* **29**, 159-183.

Rubin, D. B. (1973b). The use of matched sampling and regression adjustments to remove bias in observational studies, *Biometrics* **29**, 185-203.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology* **66**, 688-701.

Rubin, D. B. (1990a). Neyman (1923) and causal inference in experiments and observational studies, *Statistical Science* **5**, 472-480.

Rubin, D. B. (1990b). Formal modes of statistical inference for causal effects, *Journal of Statistical Planning and Inference* **25**, 279-292.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism, *Biometrics* **47**, 1213-1234.

Appendix: Sampling properties of $\widehat{\Delta}_h(t;W)$

Sampling scheme

We adapt the sampling model of Neyman (1990) to our context. Let us consider two urns representing each one of the two potential outcomes. For a given time W, urn k, $k = 0, 1$, contains the potential outcomes (possibly censored) $T_i^k(W)$ for $i = 1, \ldots, 2n_1$, the n_1 matched pairs that have survived up to time W. At a given time $t > W$, the elements in the urns are of three different types: $(\times)$ those who have died or have been censored before time t ($T_i^k(W) < t$ or censored before t, $k = 0, 1$), (0) those who have not been censored until time t and die at time t ($T_i^k(W) = t$, $k = 0, 1$) and (1) those who have survived and have not been censored up to time t ($T_i^k(W) \geq t$, $k = 0, 1$). Denote by $m_\times^k(t)$, $m_0^k(t)$, and $m_1^k(t)$, respectively, the number of elements in the three categories for the two urns $k = 0, 1$.

These two urns describe the population in the inferential framework adopted in this paper. From this population, at the beginning of the substudy conditional on waiting time W, we sample without replacement n_1 individuals from one of the urn. Each time one individual is sampled, say from the urn with the treated potential outcome, the non-treated potential outcome corresponding to the same unit is also removed from the other urn.

Assume that n_1 individuals are drawn without replacement from urn k, and define by

Y^k : the number of individuals (out of the n_1 sampled) which are of type (1);

X^k : the number of individuals among the Y^k above which are of type (0).

Then, Y^k is a hypergeometric random variable with parameters $(2n_1, n_1, \frac{m_1^k(t)}{2n_1})$. Similarly, $X^k|Y^k = y$ is also hypergeometric with parameters $(m_1^k(t), y, \frac{m_0^k(t)}{m_1^k(t)})$.

Unbiasedness

We show here the unbiasedness of $\widehat{\Delta}_h(t;W)$ for $\Delta_h(t;W)$, i.e. $E\left(\widehat{\Delta}_h(t;W)\right) = \Delta_h(t;W)$, where the expectation operator is defined by the sampling

scheme described above. Note that $\widehat{h}^k(t;W) = \frac{X^k}{Y^k}$, $k = 0, 1$. Then, using the distribution identified above we have

$$\begin{aligned} E(\widehat{h}^1(t;W)) &= E(X^1/Y^1) \\ &= E(E(X^1/Y^1|Y^1)) = E\left(\frac{1}{Y^1}E(X^1|Y^1)\right) \\ &= E\left(\frac{1}{Y^1}Y^1\frac{m_0^1(t)}{m_1^1(t)}\right) = \frac{m_0^1(t)}{m_1^1(t)} = h^1(t;W). \end{aligned}$$

Similarly, for the urn with the controls, we have

$$E(\widehat{h}^0(t;W)) = \frac{m_0^0(t)}{m_1^0(t)} = h^0(t;W).$$

Hence, $E\left(\widehat{\Delta}_h(t;W)\right) = \Delta_h(t;W)$.

Variance

We want to estimate $Var(\widehat{h}^1(t;W) - \widehat{h}^0(t;W))$. We have

$$\begin{aligned} Var(\widehat{h}^1(t;W) - \widehat{h}^0(t;W)) &= Var(X^1/Y^1 - X^0/Y^0) \\ &= E(Var(X^1/Y^1 - X^0/Y^0|Y^1, Y^0)) \\ &\quad + Var(E(X^1/Y^1 - X^0/Y^0|Y^1, Y^0)) \\ &= E(Var(X^1/Y^1 - X^0/Y^0|Y^1, Y^0)). \end{aligned}$$

The last equality follows because $E(X^1/Y^1 - X^0/Y^0|Y^1, Y^0)$ is constant; see previous section. By the same arguments used in Neyman (1990) −see also Rubin (1990a, Eq. 2)− we can write

$$\begin{aligned} &Var(X^1/Y^1 - X^0/Y^0|Y^1, Y^0) \\ &= E\left(\frac{\widehat{h}^1(t;W)(1-\widehat{h}^1(t;W))}{Y^1 - 1} + \frac{\widehat{h}^0(t;W)(1-\widehat{h}^0(t;W))}{Y^0 - 1}\middle| Y^1, Y^0\right) \\ &\qquad - \frac{1}{Y^1 + Y^0}S^2, \end{aligned}$$

where

$$
\begin{aligned}
S^2 &= \frac{1}{Y^1+Y^0-1}\sum_{i=1}^{Y^1+Y^0}\left[I(T_i^1(W)=t)-I(T_i^0(W)=t)-\overline{D}\right]^2, \\
\overline{D} &= \frac{1}{Y^1+Y^0}\sum_{i=1}^{Y^1+Y^0}\left[I(T_i^1(W)=t)-I(T_i^0(W)=t)\right].
\end{aligned}
$$

This is, in particular, due to the fact that $\widehat{h}^j(t;W)(1-\widehat{h}^j(t;W))$ is the usual variance estimator based on a sample of size Y_j, $j=0,1$. Note that for us to be able to use the results of Neyman we have to condition on Y^1 and Y^0. Putting the two previous results together we have

$$
\begin{aligned}
&Var(\widehat{h}^1(t;W)-\widehat{h}^0(t;W)) \\
&\quad = E\left(\frac{\widehat{h}^1(t;W)(1-\widehat{h}^1(t;W))}{Y^1-1}+\frac{\widehat{h}^0(t;W)(1-\widehat{h}^0(t;W))}{Y^0-1}\right) \\
&\qquad\qquad - E\left(\frac{1}{Y^1+Y^0}S^2\right).
\end{aligned}
$$

This tells us that $\frac{\widehat{h}^1(t;W)(1-\widehat{h}^1(t;W))}{Y^1-1}+\frac{\widehat{h}^0(t;W)(1-\widehat{h}^0(t;W))}{Y^0-1}$ is an unbiased estimator of $Var(\widehat{h}^1(t;W)-\widehat{h}^0(t;W))$ when S^2=0, that is, for instance, when $T_i^1(W)=T_i^0(W)$, for $i=1,\ldots,2n_1$. In general, the estimator of the variance is positively biased.

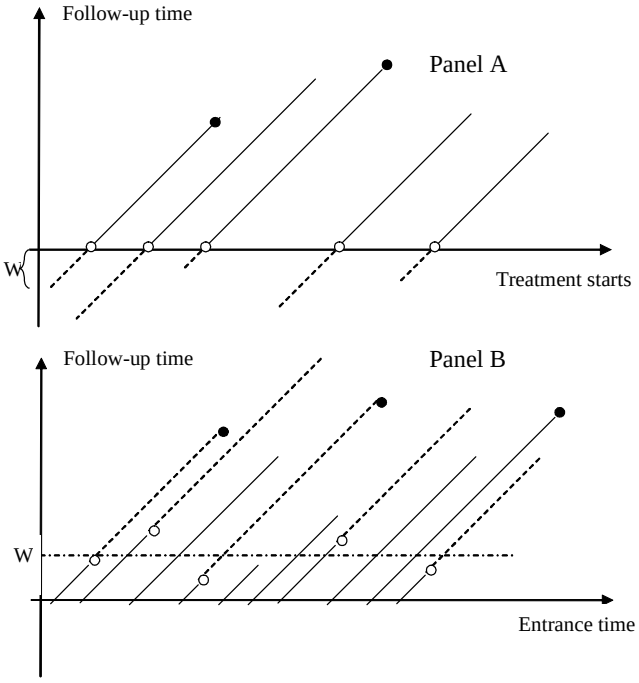

Figure 1: The x-axis represents calender time and the y-axis represents time with origin when the patients are treated (Panel A) or enter the study (Panel B). Treatments are represented by an open circle. The diagonal lines represent the history of each patient. Exits (deaths) are denoted with a filled circle, and W denotes the first treated patient waiting time.

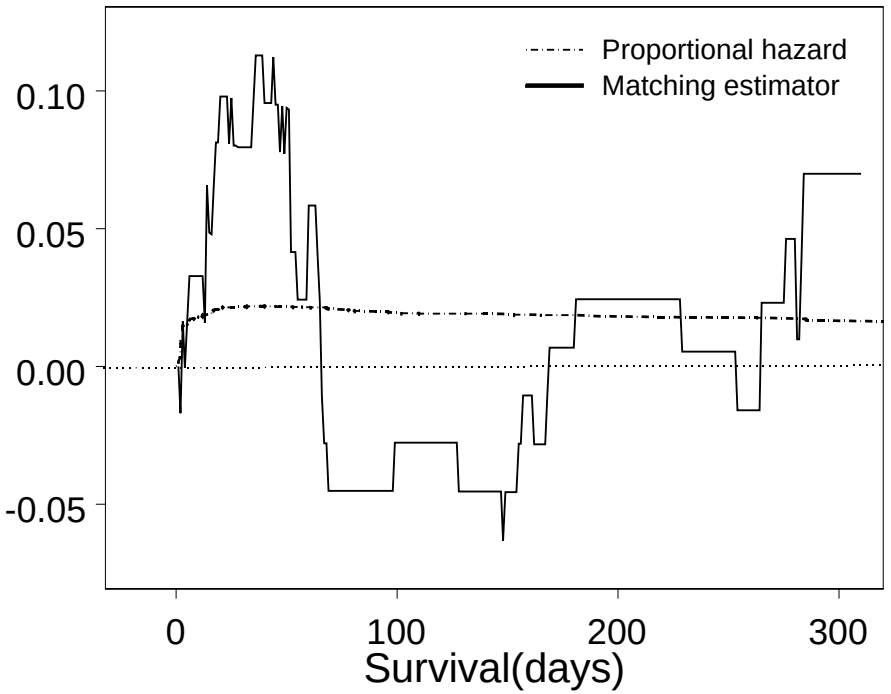

Figure 2: Estimated (by matching) treatment effect $\widehat{\overline{\Delta}}_s(t)$ (plain line) and the difference in survival functions resulting from the Cox proportional hazard model (9).

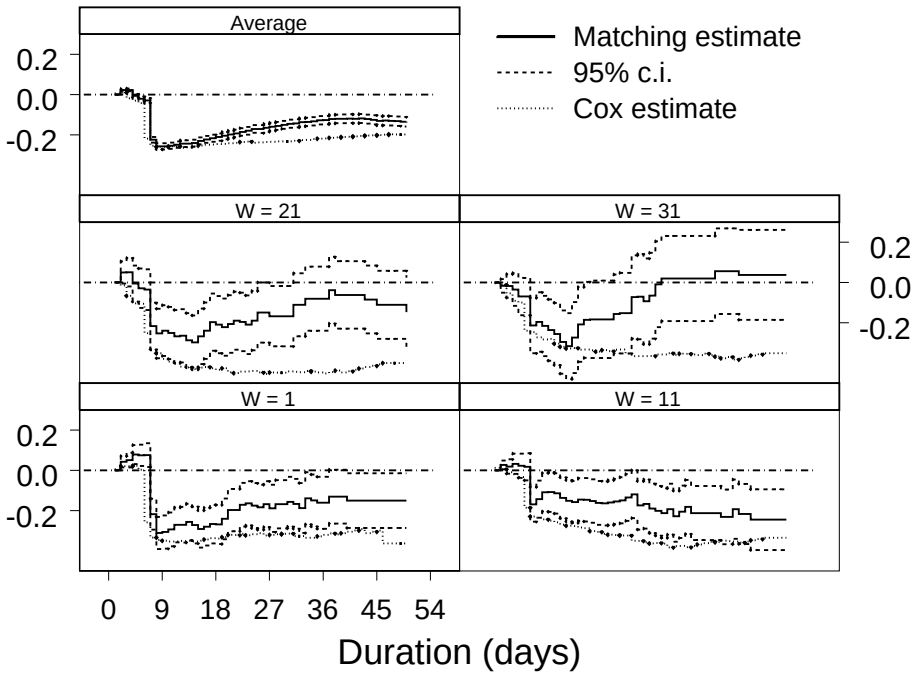

Figure 3: Estimates (including approximate 95 % confidence intervals) of the effect $\widehat{\Delta}_s(t;W)$ of employment subsidy on the duration in unemployment for $W = 1, 11, 21, 31$ and the average $\widehat{\overline{\Delta}}_s(t)$.

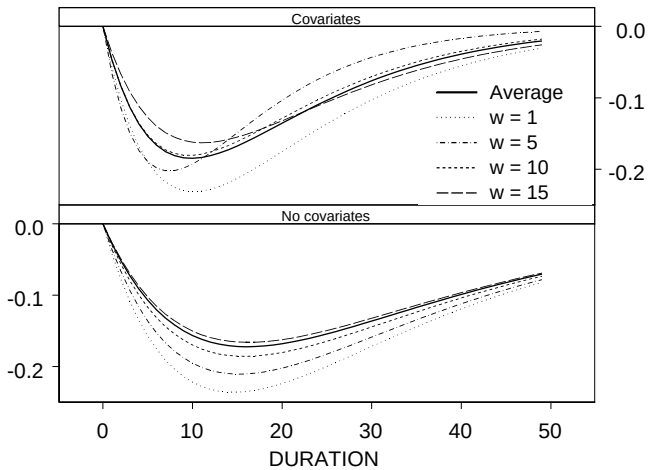

Figure 4: The treatment effects $\widetilde{\Delta}_s(t;W)$ for $W = 1, 5, 10$ and 15. The average $\overline{\widetilde{\Delta}}_s(t;W)$ is computed analytically over the observed W.

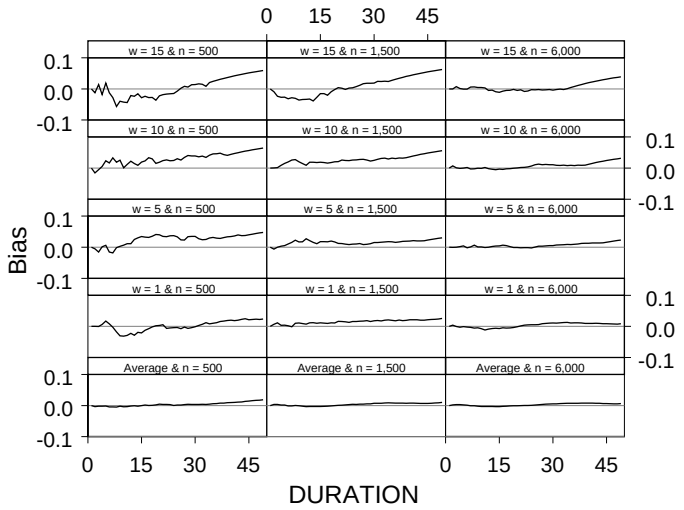

Figure 5: Bias for the matching estimator in the heterogeneous treatment setting (i.e. $a_1 = b_1 = 1$ and $a_w \neq 0$) with proportion of treated equal to 2.9%.

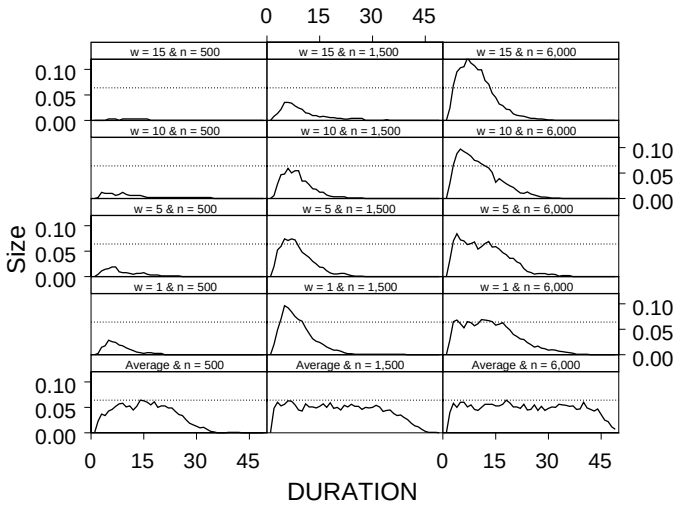

Figure 6: Empirical size (nominal size 5 %) for the matching estimator in the heterogeneous treatment setting (i.e. $a_1 = b_1 = 1$ and $a_w \neq 0$) with proportion of treated equal to 2.9%. Empirical sizes above the horizontal dotted line (6.4%) are significantly higher (at a 2.5% level of significance) than the nominal size of 5%.

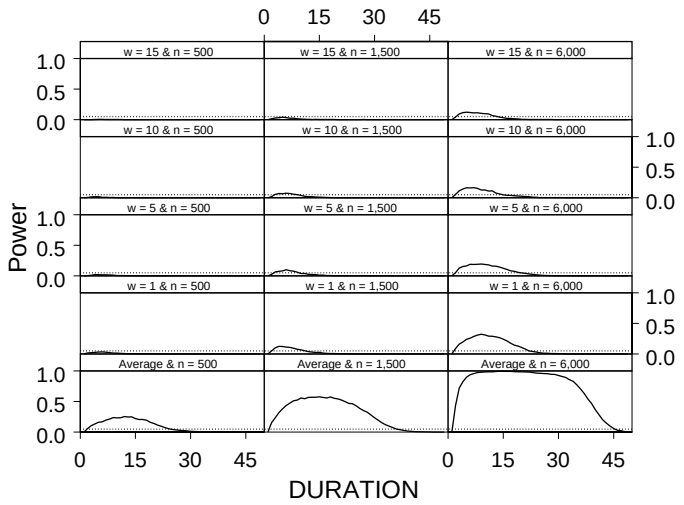

Figure 7: Power for the matching estimator in the heterogeneous treatment setting (i.e. $a_1 = b_1 = 1$and $a_w \neq 0$) with proportion of treated equal to 2.9%. The dotted line shows the nominal size 5%.

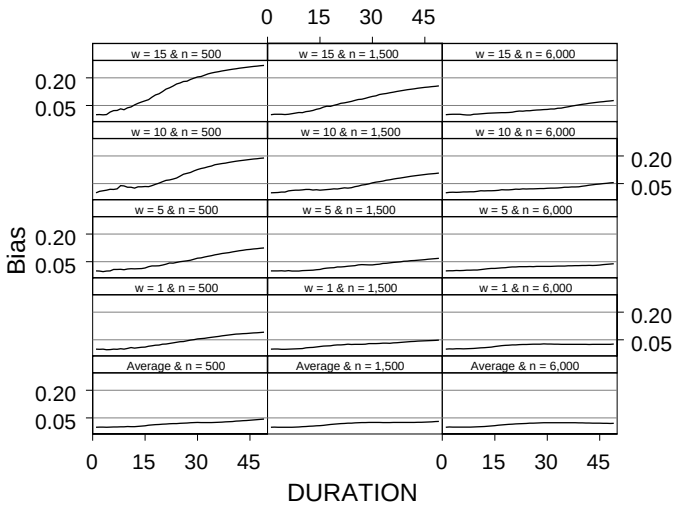

Figure 8: Bias for the matching estimator in the heterogeneous treatment setting (i.e. $a_1 = b_1 = 1$ and $a_w \neq 0$) with proportion of treated equal to 49%.

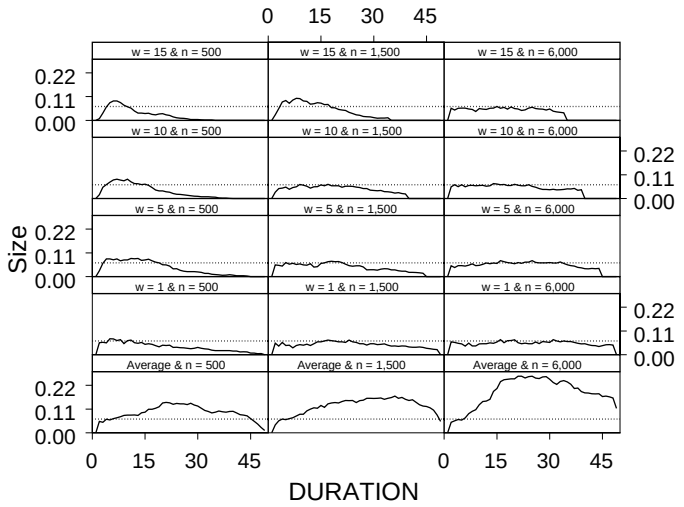

Figure 9: Empirical size (nominal size 5%) for the matching estimator in the heterogeneous treatment setting (i.e. $a_1 = b_1 = 1$ and $a_w \neq 0$) with proportion of treated equal to 49%. Empirical sizes above the horizontal dotted line (6.4%) are significantly higher (at a 2.5% level of significance) than the nominal size of 5%.

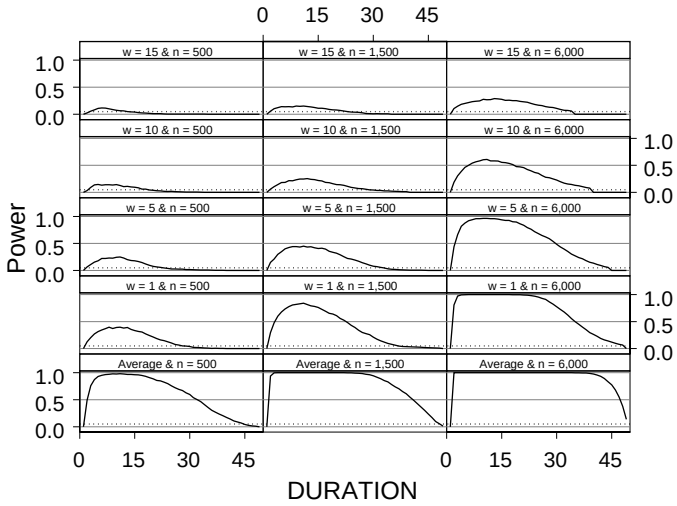

Figure 10: Power for the matching estimator in the heterogeneous treatment setting (i.e. $a_1 = b_1 = 1$ and $a_w \neq 0$) with proportion of treated equal to 49%. The dotted line shows the nominal size 5%.

Publication series published by the Institute for Labour Market Policy Evaluation (IFAU) – latest issues

Rapporter/Reports

2006:1 Zenou Yves, Olof Åslund & John Östh ”Hur viktig är närheten till jobb för chanserna på arbetsmarknaden?”

2006:2 Mörk Eva, Linus Lindqvist & Daniela Lundin ”Påverkar maxtaxan inom barnomsorgen hur mycket föräldrar arbetar?”

2006:3 Hägglund Pathric ”Anvisningseffekter” – finns dom? Resultat från tre arbetsmarknadspolitiska experiment”

2006:4 Hägglund Pathric ”A description of three randomised experiments in Swedish labour market policy”

2006:5 Forslund Anders & Oskar Nordström Skans “(Hur) hjälps ungdomar av arbetsmarknadspolitiska program för unga?”

2006:6 Johansson Per & Olof Åslund ”’Arbetsplatsintroduktion för vissa invandrare’ – teori, praktik och effekter”

2006:7 Calleman Catharina ”Regleringen av arbetsmarknad och anställningsförhållanden för hushållstjänster”

2006:8 Nordström Skans Oskar, Per-Anders Edin & Bertil Holmlund ”Löneskillnader i svenskt näringsliv 1985–2000”

2006:9 Engström Per, Hesselius Patrik & Malin Persson ”Överutnyttjande i tillfällig föräldrapenning för vård av barn”

2006:10 Holmlund Bertil, Qian Liu & Oskar Nordström Skans ”Utbildning nu eller senare? Inkomsteffekter av uppskjuten högskoleutbildning”

2006:11 Sibbmark Kristina & Olof Åslund ”Vad för vem och hur gick den sen? En kartläggning av arbetsförmedlingarnas insatser för utrikes födda under 2005”

2006:12 Fredriksson Peter & Björn Öckert ”Är det bättre att börja skolan tidigare?”

2006:13 Sibbmark Kristina ”Arbetsmarknadspolitisk översikt 2005”

2006:14 Wang Iris J Y, Kenneth Carling & Ola Nääs ”Är kommunala sommarjobb en gräddfil till arbetsmarknaden?”

2006:15 Söderström Martin ”Betygsintagning och elevers studieframgång i Stockholms gymnasieskolor”

2006:16 Johansson Kerstin **”**Arbetskraftsdeltagande och arbetsmarknadspolitiska program”

2007:1 Lundin Daniela ”Subventionerade anställningar för unga – en uppföljning av allmänt anställningsstöd för 20–24-åringar”

Working Papers

2006:1 Åslund Olof, John Östh & Yves Zenou "How important is access to jobs? Old question – improved answer"

2006:2 Hägglund Pathric "Are there pre-programme effects of Swedish active labour market policies? Evidence from three randomised experiments"

2006:3 Johansson Per "Using internal replication to establish a treatment effect"

2006:4 Edin Per-Anders & Jonas Lagerström "Blind dates: quasi-experimental evidence on discrimination"

2006:5 Öster Anna "Parental unemployment and children's school performance"

2006:6 Forslund Anders & Oskar Nordström Skans "Swedish youth labour market policies revisited"

2006:7 Åslund Olof & Per Johansson "Virtues of SIN – effects of an immigrant workplace introduction program"

2006:8 Lalive Rafael "How do extended benefits affect unemployment duration? A regression discontinuity approach"

2006:9 Nordström Skans Oskar, Per-Anders Edin & Bertil Holmlund "Wage dispersion between and within plants: Sweden 1985–2000"

2006:10 Korkeamäki Ossi & Roope Uusitalo "Employment effects of a payroll-tax cut – evidence from a regional tax exemption experiment"

2006:11 Holmlund Bertil, Qian Liu & Oskar Nordström Skans "Mind the gap? Estimating the effects of postponing higher education"

2006:12 Fredriksson Peter & Björn Öckert "Is early learning really more productive? The effect of school starting age on school and labor market performance"

2006:13 Pekkarinen Tuomas, Sari Pekkala & Roope Uusitalo "Educational policy and intergenerational income mobility: evidence from the Finnish comprehensive school reform"

2006:14 Wang Iris J Y, Kenneth Carling & Ola Nääs "High school students' summer jobs and their ensuing labour market achievement"

2006:15 de Jong Philip, Maarten Lindeboom & Bas van der Klaauw "Screening disability insurance applications"

2006:16 Söderström Martin "School choice and student achievement – new evidence on open-enrolment"

2006:17 Johansson Kerstin "Do labor market flows affect labor-force participation?"

2007:1 de Luna Xavier & Per Johansson "Matching estimators for the effect of a treatment on survival times"

Dissertation Series

2006:1 Hägglund Pathric "Natural and classical experiments in Swedish labour market policy"

2006:2 Savvidou Eleni "Technology, human capital and labor demand"

2006:3 Söderström Martin "Evaluating institutional changes in education and wage policy"

2006:4 Lagerström Jonas "Discrimination, sickness absence, and labor market policy"

2006:5 Johansson Kerstin " Empirical essays on labor-force participation, matching, and trade"